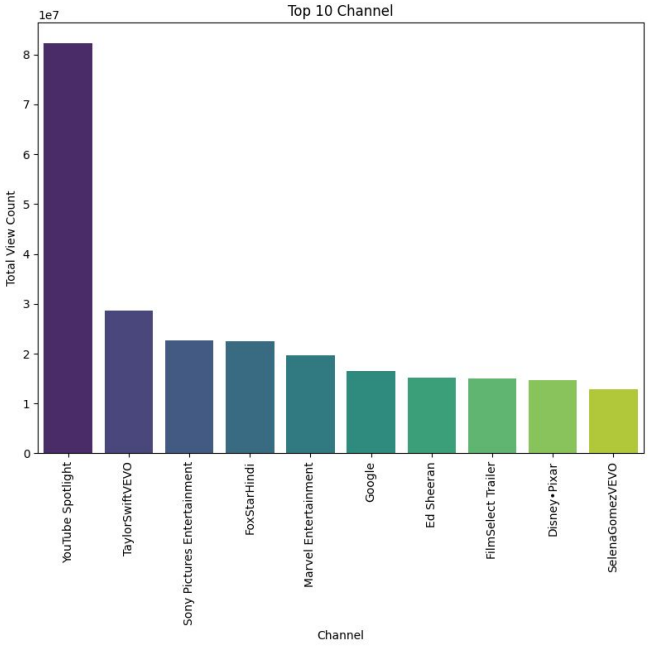
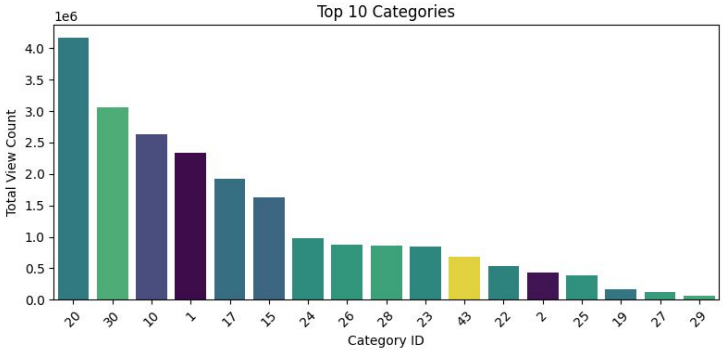
# Youtube Views Prediction

Agi Rahmawandi
Batch 57

# Exploratory Data Analysis




Top 5 Youtube Videos

| | title | views | likes |
|---|---|---|---|
| YouTube Rewind: The Shape of 2017 \| #YouTubeRe... | 125432237 | 2912710 |
| YouTube Rewind: The Shape of 2017 \| #YouTubeRe... | 113876217 | 2811216 |
| YouTube Rewind: The Shape of 2017 \| #YouTubeRe... | 100911567 | 2656672 |
| Marvel Studios' Avengers: Infinity War Officia... | 89930713 | 2606663 |
| Marvel Studios' Avengers: Infinity War Officia... | 87449453 | 2584674 |

# Exploratory Data Analysis

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36791 entries, 0 to 36790
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   trending_date           36791 non-null  object
 1   title                   36791 non-null  object
 2   channel_title           36791 non-null  object
 3   category_id             36791 non-null  int64
 4   publish_time            36791 non-null  object
 5   tags                    36791 non-null  object
 6   views                   36791 non-null  int64
 7   likes                   36791 non-null  int64
 8   dislikes                36791 non-null  int64
 9   comment_count           36791 non-null  int64
 10  comments_disabled       36791 non-null  bool
 11  ratings_disabled        36791 non-null  bool
 12  video_error_or_removed  36791 non-null  bool
 13  description             36746 non-null  object
 14  No_tags                 36791 non-null  int64
 15  desc_len                36791 non-null  int64
 16  len_title               36791 non-null  int64
 17  publish_date            36791 non-null  datetime64[ns]
dtypes: bool(3), datetime64[ns](1), int64(8), object(6)
memory usage: 4.3+ MB
```

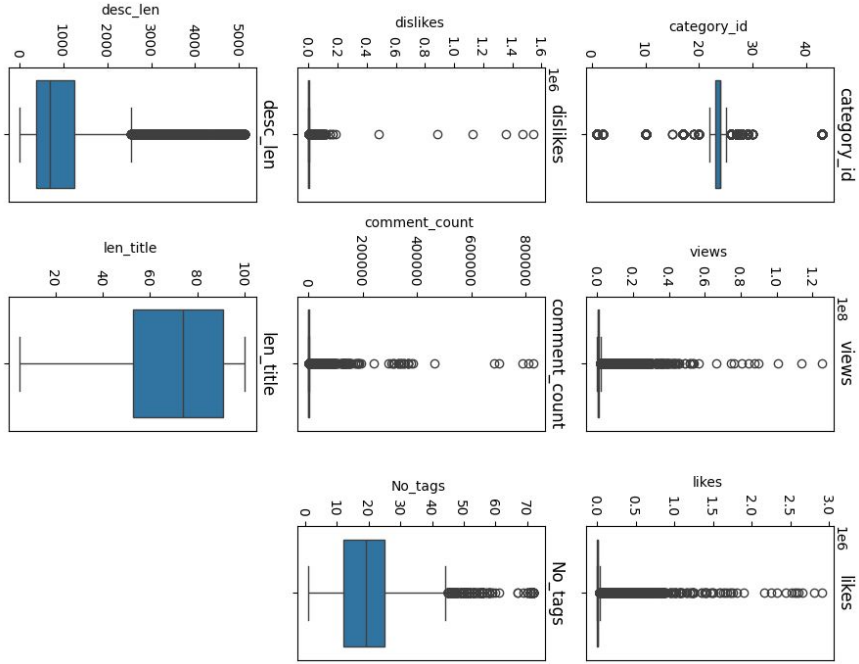
- Terdapat nilai null pada kolom description
- Type data tidak sesuai pada colom trending_date dan publish_time harusnya type data datetime

# Exploratory Data Analysis

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| category_id | 36791.00 | 21.55 | 6.59 | 1.00 | 23.00 | 24.00 | 24.00 | 43.00 |
| views | 36791.00 | 1071490.26 | 3207149.05 | 4024.00 | 125604.00 | 307836.00 | 806631.50 | 125432237.00 |
| likes | 36791.00 | 27450.69 | 97831.29 | 0.00 | 879.00 | 3126.00 | 14095.00 | 2912710.00 |
| dislikes | 36791.00 | 1685.36 | 16197.32 | 0.00 | 109.00 | 331.00 | 1032.00 | 1545017.00 |
| comment_count | 36791.00 | 2714.02 | 14978.11 | 0.00 | 83.00 | 336.00 | 1314.50 | 827755.00 |
| No_tags | 36791.00 | 18.94 | 9.84 | 1.00 | 12.00 | 19.00 | 25.00 | 72.00 |
| desc_len | 36791.00 | 923.08 | 815.04 | 3.00 | 368.00 | 677.00 | 1237.00 | 5136.00 |
| len_title | 36791.00 | 70.61 | 22.41 | 5.00 | 53.00 | 74.00 | 91.00 | 100.00 |

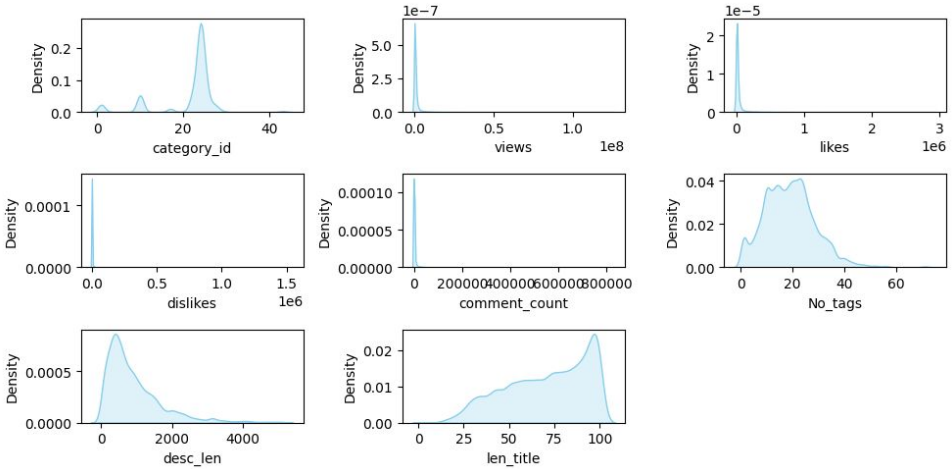
- Dari nilai standar deviasinya data cukup beragam
- Pada beberapa kolom terdapat perbedaan yang cukup jauh antara mean dan mediannya artinya data memiliki outlier

# Exploratory Data Analysis



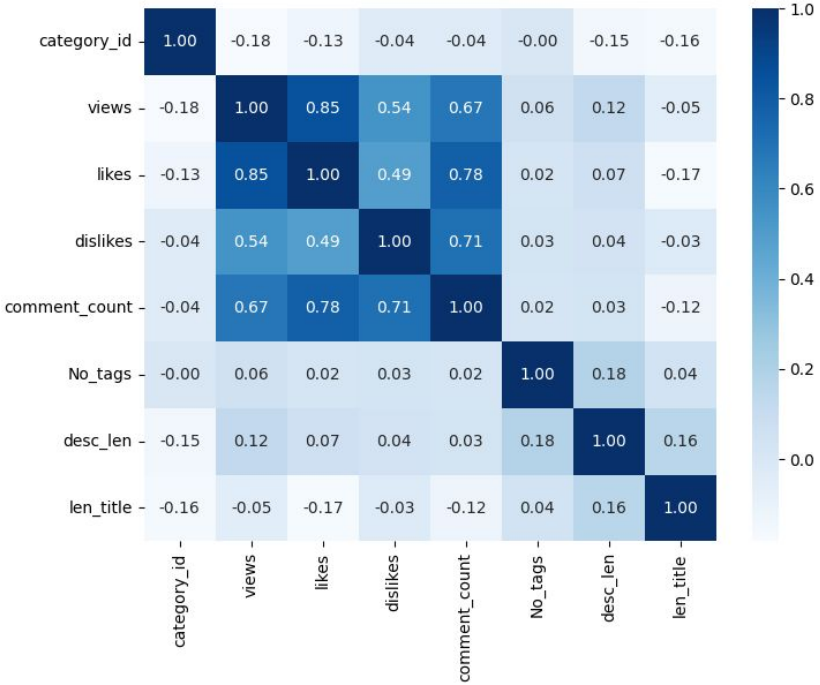

- Hampir semua kolom numerik terihat ada outlier terutapa pada kolom views dislike, like dan comment count terdapat outlier yang cukup jauh dari iqrnya

# Exploratory Data Analysis




- hampir seluruh kolom datanya tidak terdistribusi secara normal/skewed

# Exploratory Data Analysis




- Kolom yang berkorelasi kuat dengan target (‘views’) ialah like, dislike,comment count berada di atas 0.5.
- Sama halnya korelasi antar feature ketiga kolom like, dislike, serta commen_count berkorelasi cukup kuat(redundan)
- Category_id, no_tags, desc_len, len_title memiliki korelasi hampi mendekati nol dengan sebagian besar feature.

# Exploratory Data Analysis

Dari hasil EDA dapat ditarik kesimpulan bahwa :

- Akan dilakukan penyesuaian pada kolom description yang memiliki nilai null
- Menyesuiakan tipe data pada kolom trending_date dan publish_time
- Pembersihan outlier pada kolom-kolom yang memiliki outlier
- Feature yang akan digunakan semua kolom numeric, kecuali kolom comment_count akan di drop/dihilanngkan karena redundan dengan kolom like dan dislike

# Data Cleansing

- Drop baris bernilai null pada kolom description
- Permbesihan data dengan mempertahankan data yang pertama muncul dengan `duplicated(keep='first')`terdapat data duplikat sebesar 4228 dan akan dihapus.
- Penyesuaian type data pada kolom trending_date dan publish_time menjadi datetime
- Handling outlier dengan IQR

# Feature Engineering

**Feature Selection**

- Semua table numeric bertipe float,int, dan boolean dipilih menjadi feature kecuali kolom comment_count
- dislike, like, comment_count ketiganya memiliki nilai korelasi yang kuat satu sama lain, dan berkorelasi kuat juga dengan target, maka salah satu kolom akan didrop yakni kolom comment_count karena redundan cukup tinggi dengan kolom like dan dislike.

# Feature Engineering

**Feature encoding**

- Karena kolom category_id berjenis kategori maka dilakukan encoding dengan one hot encoding.

**Feature encoding**

- Membuat feature baru dari feature publish_date untuk menjadi kolom hari, yang tujuannya melihat pada hari apa, publish video yang berpotensi mendapatkan jumlah views banyak.

# Modeling

- Memilih feature numeric dan boolean
- Membagi data train dan data test sebersar ⅓ untuk data test dan sisanya untuk data train.
- Standarisasi feature menggunkan menggunkan robustscaller, karena meruapakan teknik penskalaan data yang tidak sensitif terhadap outlier,teknik ini menggunakan median dan interquartile range (IQR).
- Standarisasi dilakukan hanya pada feature.
- Model yang digunakan adalah LinearRegression,RandomForest, HistGradientBoosting, dan LGBMRegressor.

# Model Evaluation

| Model | **Linear Regression** | | |
|---|---|---|---|
| | **Training Metrics** | | Underfitting |
| | R^2 | 0.468 | |
| | MAE | 143352.6074 | |
| | RMSE | 213248.7381 | |
| | | | |
| | **Testing Metrics** | | |
| | R^2 | 0.4737 | |
| | MAE | 143459.0941 | |
| | RMSE | 214821.6837 | |

| Model | **RandomForest** | | |
|---|---|---|---|
| | **Training Metrics** | | Overfitting |
| | R^2 | 0.9592 | |
| | MAE | 35738.8006 | |
| | RMSE | 59068.5475 | |
| | | | |
| | **Testing Metrics** | | |
| | R^2 | 0.7301 | |
| | MAE | 95618.6604 | |
| | RMSE | 153829.7892 | |

# Model Evaluation

| Model | HistGradientBoosting | | |
|---|---|---|---|
| | **Training Metrics** | | Good Fit |
| | R^2 | 0.7881 | |
| | MAE | 87826.1877 | |
| | RMSE | 134597.1508 | |
| | | | |
| | **Testing Metrics** | | |
| | R^2 | 0.6862 | |
| | MAE | 104935.3179 | |
| | RMSE | 165877.2464 | |

| Model | LGBMRegressor | | |
|---|---|---|---|
| | **Training Metrics** | | Good Fit |
| | R^2 | 0.7901 | |
| | MAE | 87686.4751 | |
| | RMSE | 133961.8174 | |
| | | | |
| | **Testing Metrics** | | |
| | R^2 | 0.6907 | |
| | MAE | 104460.5965 | |
| | RMSE | 164692.0312 | |

# Model Evaluation

- **Linear Regression**

Memiliki R^2 yang cenderung rendah yakni **0,46** pada data train dan **0,47** pada data testing, tergolong rendah, serta memiliki nilai MAE dan RMSE cukup besar **143ribu** dan **214ribu,** artinya nilai errornya cukup besar, dan data tergolong underfitting.

- **Random Forest**

Model sangat akurat di data training namun mengalami penurunan didata tetsting,begitupun nilai error mengalami lonjakan pada data testing, maka model ini tergolong overfitting,

- **HistGradientBoosting**

Dilihat di R^2 antara data tenting dan data training model cukup konsisten dan stabil, selilih nilai errornya pun masih wajar, model ini termasuk goodfit.

- **LightGBM (LGBMRegressor)**

Kondisinya sama seperti HistGradientBoosting, hanya lebih tinggi nilainya, perbedaan nilai pada datatest dan data train tergolong stabil, maka model ini termasuk goodfit

# Model Evaluation

Model yang dipilih ialah **LightGBM (LGBMRegressor)** karena :

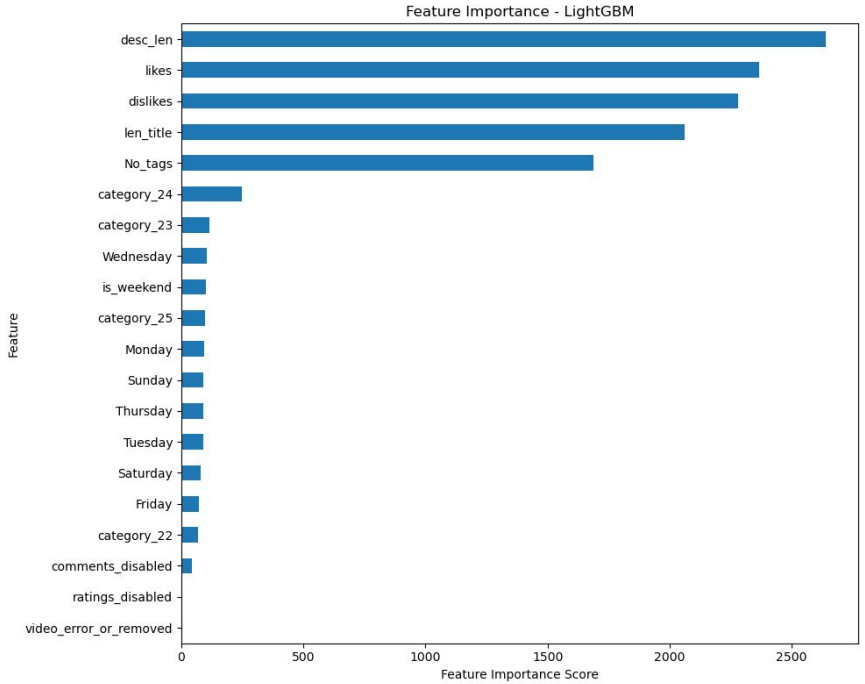
- Memilik nilai R^2 yang konsisten/stabil pada data train **0,79**, data testing **0,69**
- Memiliki nilai error paling rendah dari pada yang lain, Data train **MAE 87686 RMSE 133961,** Pada Data testing **MAE 104460 RMSE 164692.**
- Data tidak termasuk underfitting atau overfitting
- Selajutnya dilakukan tuning hyperparameter

# Tunning Hyperparameter

- Dilakukan tuning dengan metode **RandomizedSearchCV**
- Setelah melakukan tuning didapat nilai R^2 naik dari sebelum dilakukan tuning  pada data train sebesar **0,75** dan data testing sebesar **0,76**
- Hasil akhir dari model menggunakan tuning hyperparameter dari **LightGBM (LGBMRegressor)** dengan nilai hyperparameter terbaik:
    - 'subsample': 0.8,
    - 'reg_lambda': 0,
    - 'reg_alpha': 1.0,
    - 'n_estimators': 411,
    - 'max_depth': -1,
    - 'learning_rate': 0.2,
    - 'colsample_bytree': 1.0

# Feature Importance

Feature yang berperan penting dalam melakukan prediksi terhadap view



# Insight

- video dengan deskripsi panjang cenderung mendapat lebih banyak views (mungkin karena SEO, atau informasi lebih lengkap).
- Judul yang informatif atau menarik kemungkinan mengundang lebih banyak views.
- Kemungkinan tag membantu sistem rekomendasi YouTube danmengundang lebih banyak views.
- Likes/dislikes bisa jadi indikator kualitas sebuah video
- Topik popoler lebih banyak pada kategori kategori 24 dan 23.